

COMPUTING @ FERMILAB

Sharan Kalwani

Wednesday, July 01, 2015

<http://ed.fnal.gov/interns/lectures/>

<https://www.facebook.com/fermilabsist/>

COMPUTING @ FERMILAB

Developing and supporting innovative and cutting edge computing solutions and services for Fermilab



CORE COMPUTING SECTOR



Service Desk



Computer Security



E - Communication

SCIENCE & COMPUTING

G-2



CMS



Holometer



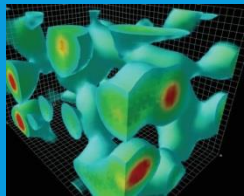
CDMS



MINOS



Lattice QCD



FTBF



NOvA



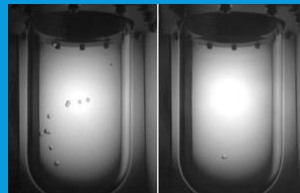
DES



Mu2e



COUPP



Pierre Auger

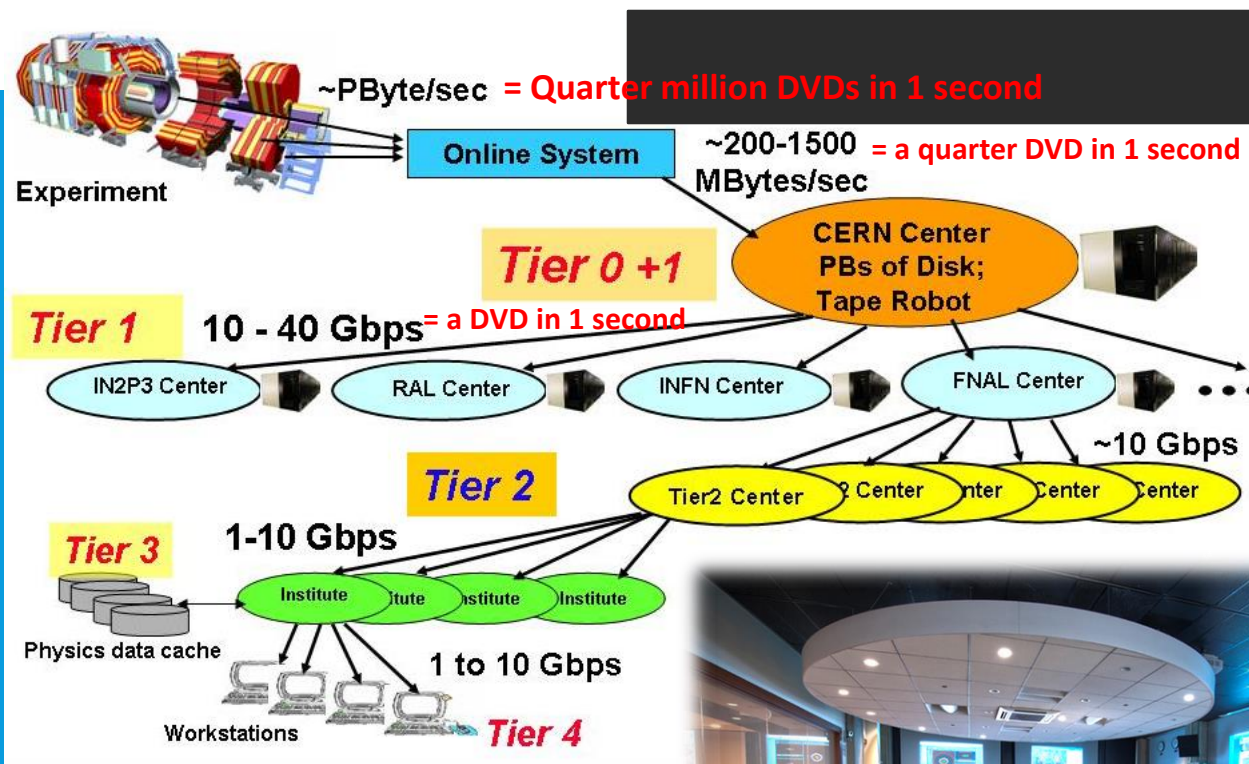


A major part of the Computing Sector's work is to support & improve the scientific programs at the lab. This includes computer support for experiment systems, design and implementation of the Data Acquisition and control systems, accelerator & detector simulations, research & development of the physics analysis software.

**Others: SeaQuest, DarkSide, MinerVa, etc,
In progress: LBNF, DUNE, LaArIAT, LARP, uBooNE,**

LHC CMS EXPERIMENT

ONE OF MANY EXPERIMENTS SUPPORTED BY THE FERMILAB COMPUTING SECTOR



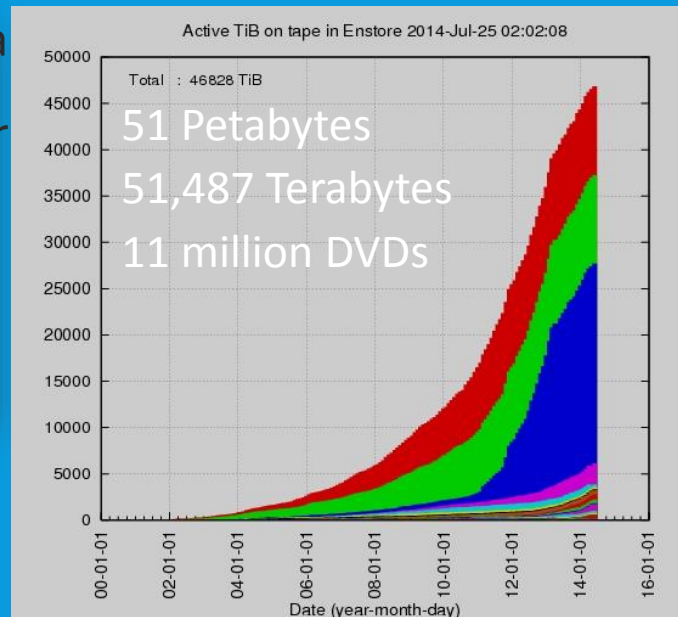
The distributed computing system for LHC distributes hundreds of Terabytes a day outward from CERN to the 11 distributed Tier-1 sites (one at Fermilab) & then 25 University Tier-2 and access to more than 100 smaller Tier-3 centers.

The CMS Remote Operations Center at Fermilab remotely supports the CMS experiment located 4,000 miles away at Cessy in France. The ROC allows US physicists to help operate the CMS detector.



DATA HANDLING & STORAGE

- **Enstore** (archival tape storage, 51 Petabytes stored, users transfer 3.2PB or 680,000 DVDs worth of data per day !!)
- **dCache** (100's of terabytes of disk front-end to Enstore for faster access)

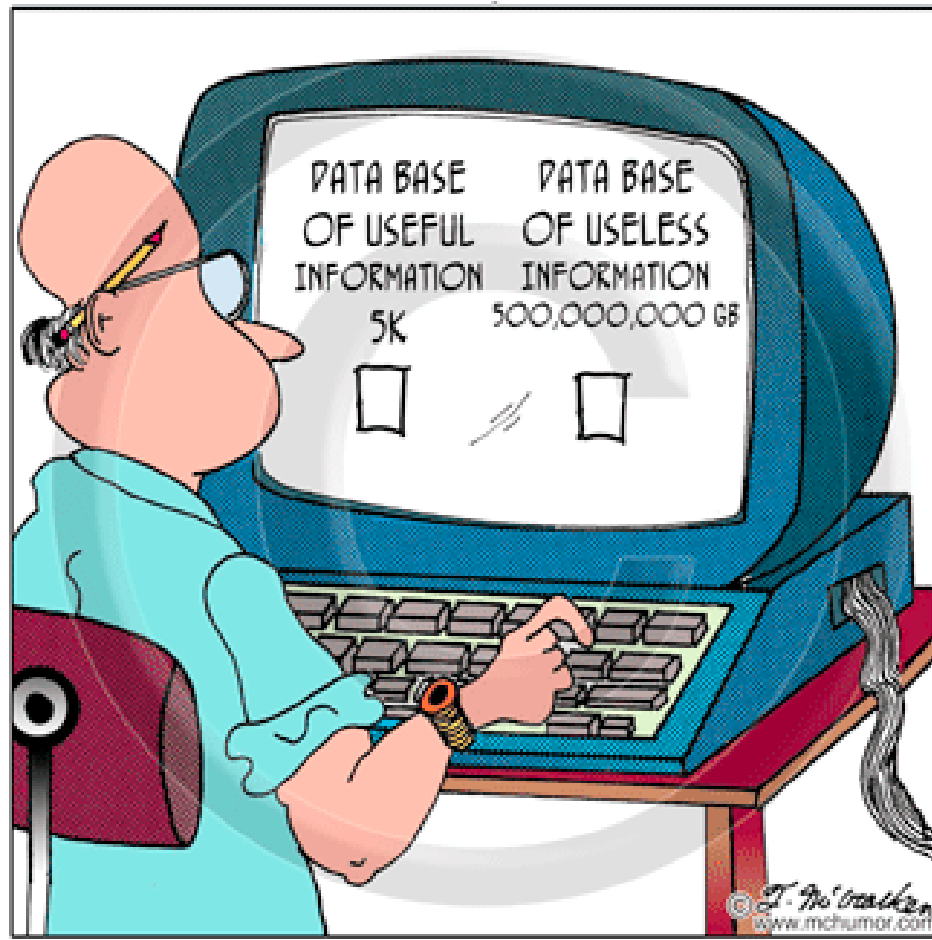


T10KC Tape, storage capacity 5TB

WHAT IS BIG DATA?

DATA BASE
OF USEFUL
INFORMATION

5K



DATA BASE
OF USELESS
INFORMATION

500,000,000 GB

.....BIG DATA?

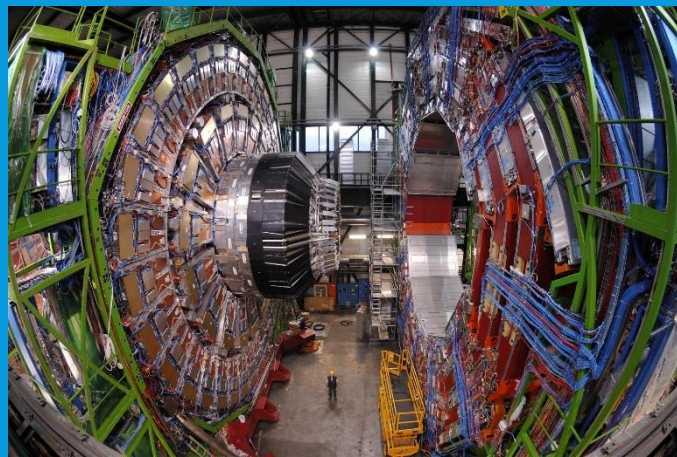
Big data is any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

Volume

Velocity

Variety

Veracity



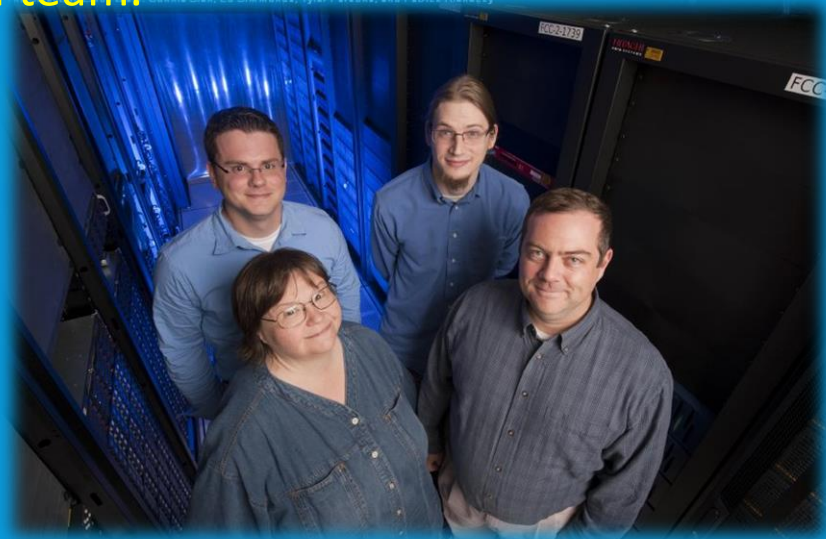
600 million collisions per second. ~150 million Petabytes per year.

100 collisions per second of interest. 0.001% of “interest” data is equal to 30 Petabytes of data per year.

CERN uses the WLCG (Worldwide LHC Computing Grid) to crunch all of the data.

WHAT IS SCIENTIFIC LINUX?

- Created in 2004 at Fermilab, Scientific Linux is a Linux operating system distribution assembled by Fermilab and CERN in collaboration with other HEP institutions.
- 100% open source and free.
- Scientific Linux is used as the computing platform for major research projects all around the globe.
- Supported by an active user community.
- Packaged by a dedicated and professional team.



SWITCHING GEARS

Computing



*High
Performance
Computing*

Any questions thus far ?

WHAT IS HPC?

High Performance Computing (HPC) uses supercomputers* and compute clusters to solve advanced computation problems



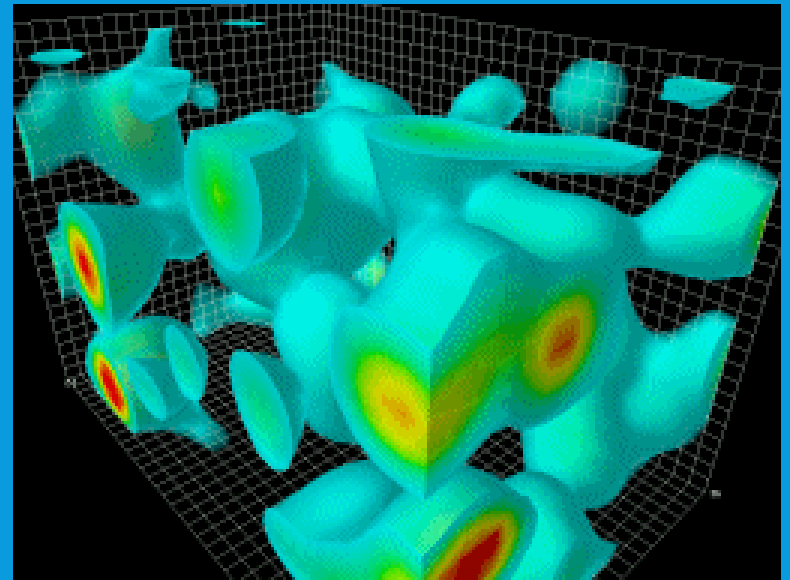
Cluster



IBM Blue Gene Supercomputer

WHY WE NEED HPC?

Proposed in the early 1970s, the theory of Quantum chromodynamics (QCD) consists of equations that describe the strong force that causes quarks to clump together to form protons and other constituents of matter. For a long time solving these equations was a struggle. But in the last decade using powerful supercomputers theorists are now able to finally solve the equations of QCD with high precision.



HOW DO I MEASURE THE SPEED OF A SUPERCOMPUTER?

FLOPS

FLoating point Operations Per Second

Examples of floating point numbers are

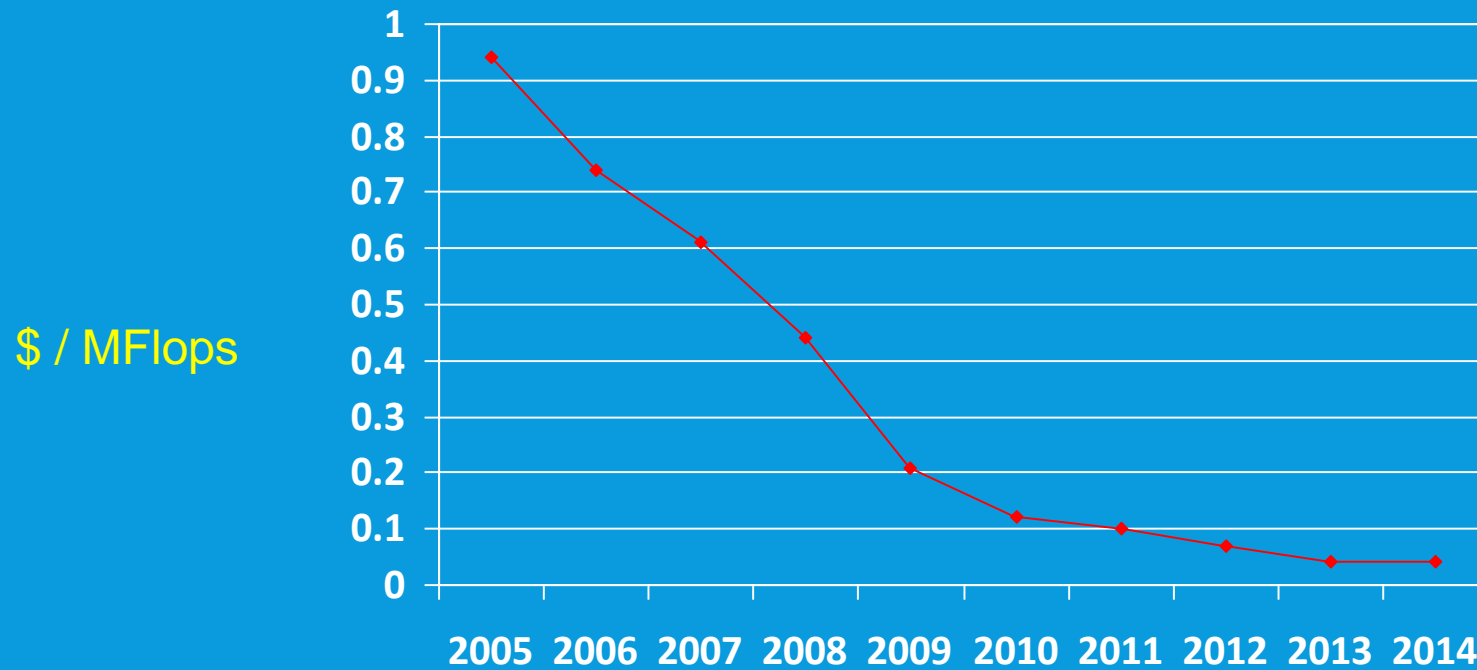
1.234567, 123456.7, 0.00001234567, 12345670000000000

LINPACK Benchmark



[\(http://www.top500.org/project/linpack/\)](http://www.top500.org/project/linpack/)

HPC COST TRENDS




How much does 1 Million Flops cost?

HPC COST TRENDS

Tracking large-scale simulation progress: Gordon Bell Prize: “price performance”

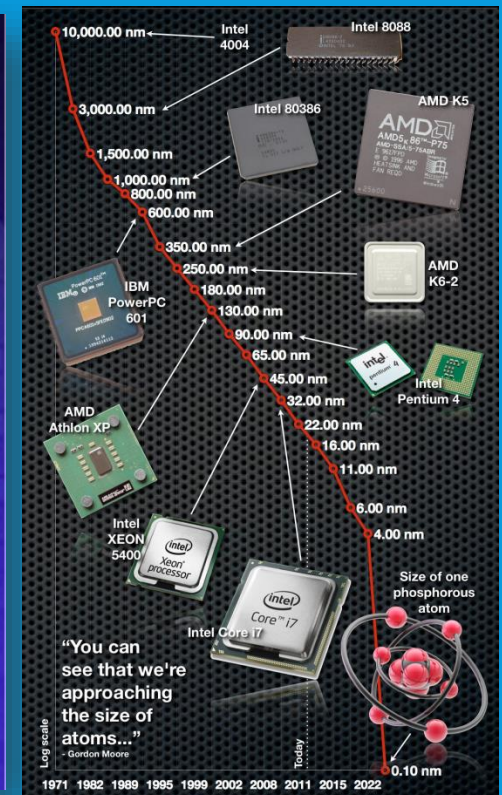
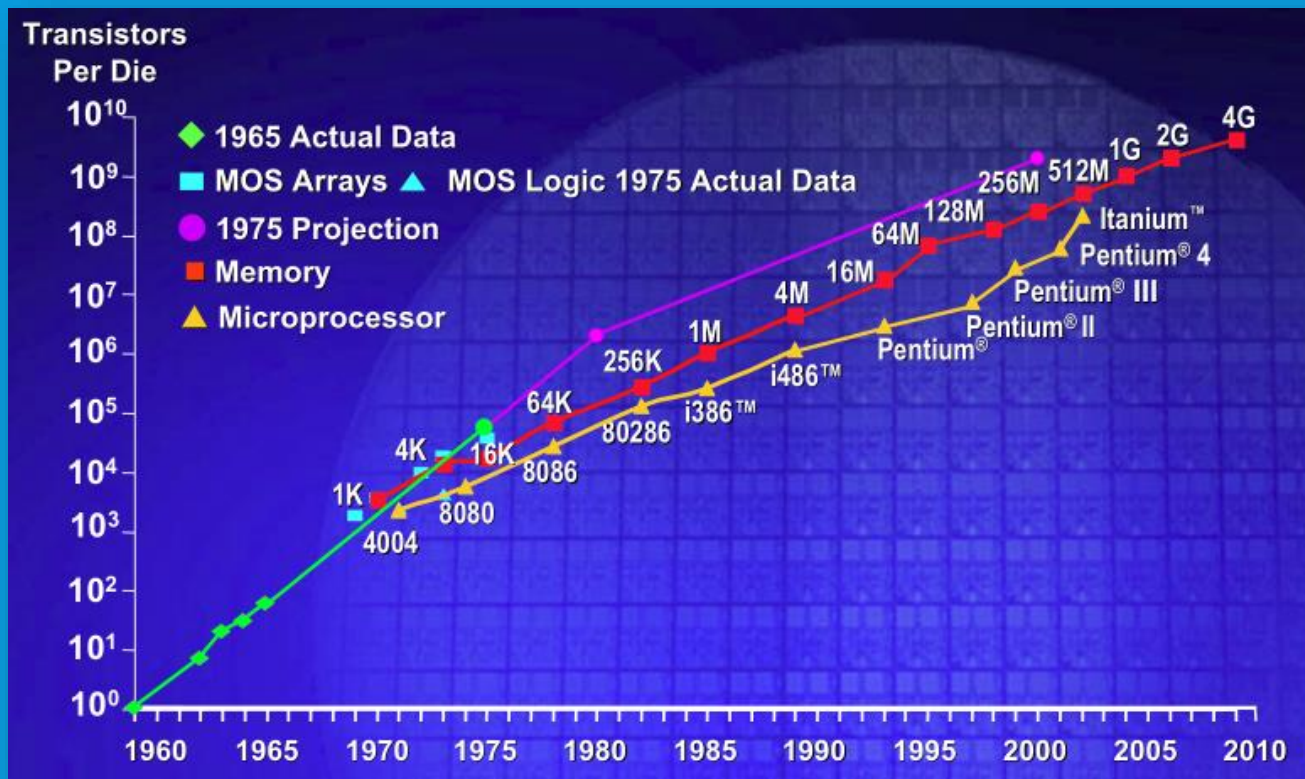
<i>Year</i>	<i>Application</i>	<i>System</i>	<i>\$ per Mflops</i>
1989	Reservoir modeling	CM-2	2,500
1990	Electronic structure	IPSC	1,250
1992	Polymer dynamics	cluster	1,000
1993	Image analysis	custom	154
1994	Quant molecular dyn	cluster	333
1995	Comp fluid dynamics	cluster	278
1996	Electronic structure	SGI	159
1997	Gravitation	cluster	56
1998	Quant chromodyn	custom	12.5
1999	Gravitation	custom	6.9
2000	Comp fluid dynamics	cluster	1.9
2001	Structural analysis	cluster	0.24
2009	Gravitation & turbulence cluster (GPU)		0.0081



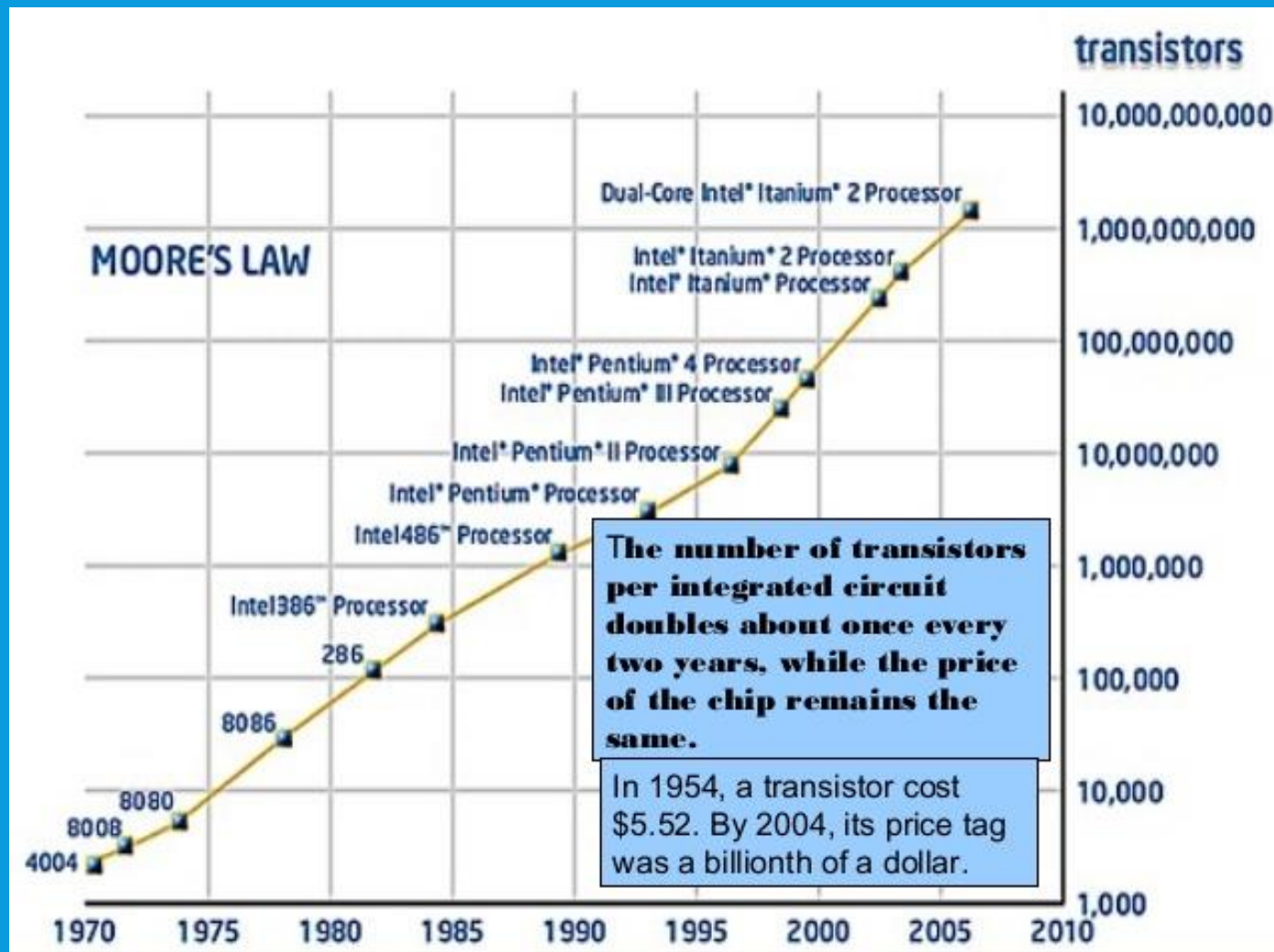
5.5 orders of
magnitude in
20 years

HPC INDUSTRY LAWS – MOORE'S LAW

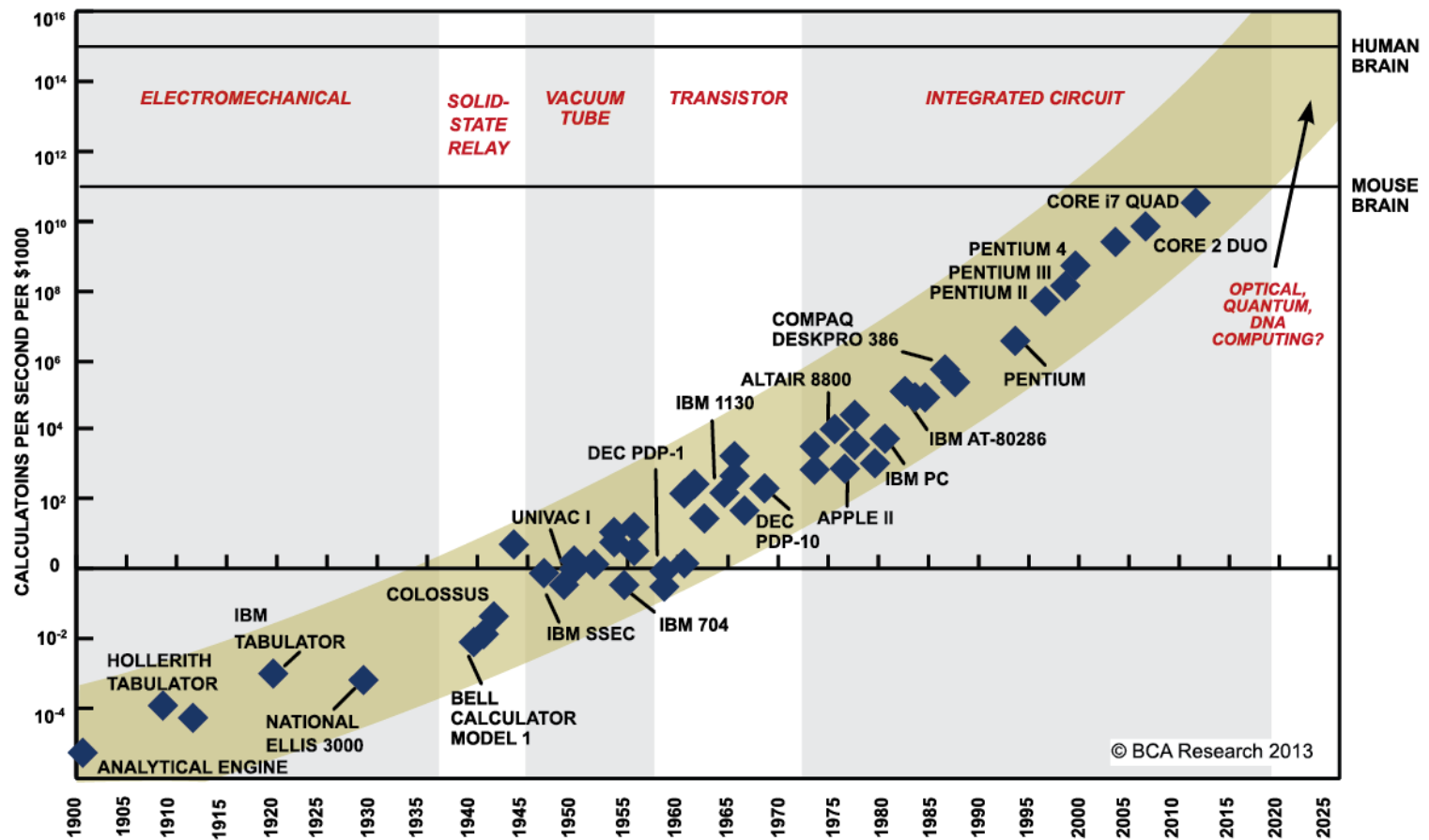
“The number of transistors that can be fabricated on a very large-scale integrated (VLSI) chip doubles every two years.” - *Intel co-founder Gordon Moore* 1965



HPC INDUSTRY LAWS – MOORE'S LAW



HPC INDUSTRY LAWS – MOORE'S LAW

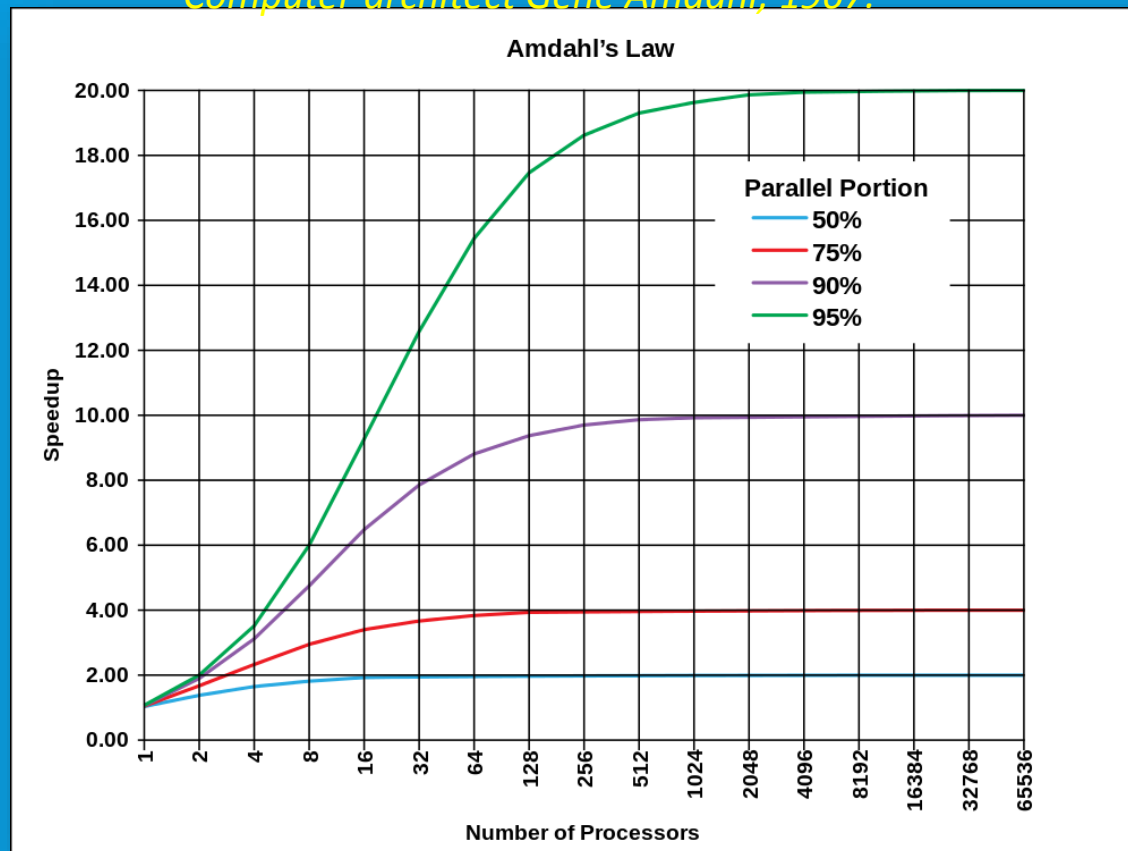


SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, THE VIKING PRESS, 2006. DATAPPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

HPC INDUSTRY LAWS – AMDAHL'S LAW

“The speedup of a program using multiple processors in parallel computing is limited by the time needed for the sequential fraction of the program.”

– Computer architect Gene Amdahl, 1967.



COMPUTE CLUSTER ARCHITECTURE

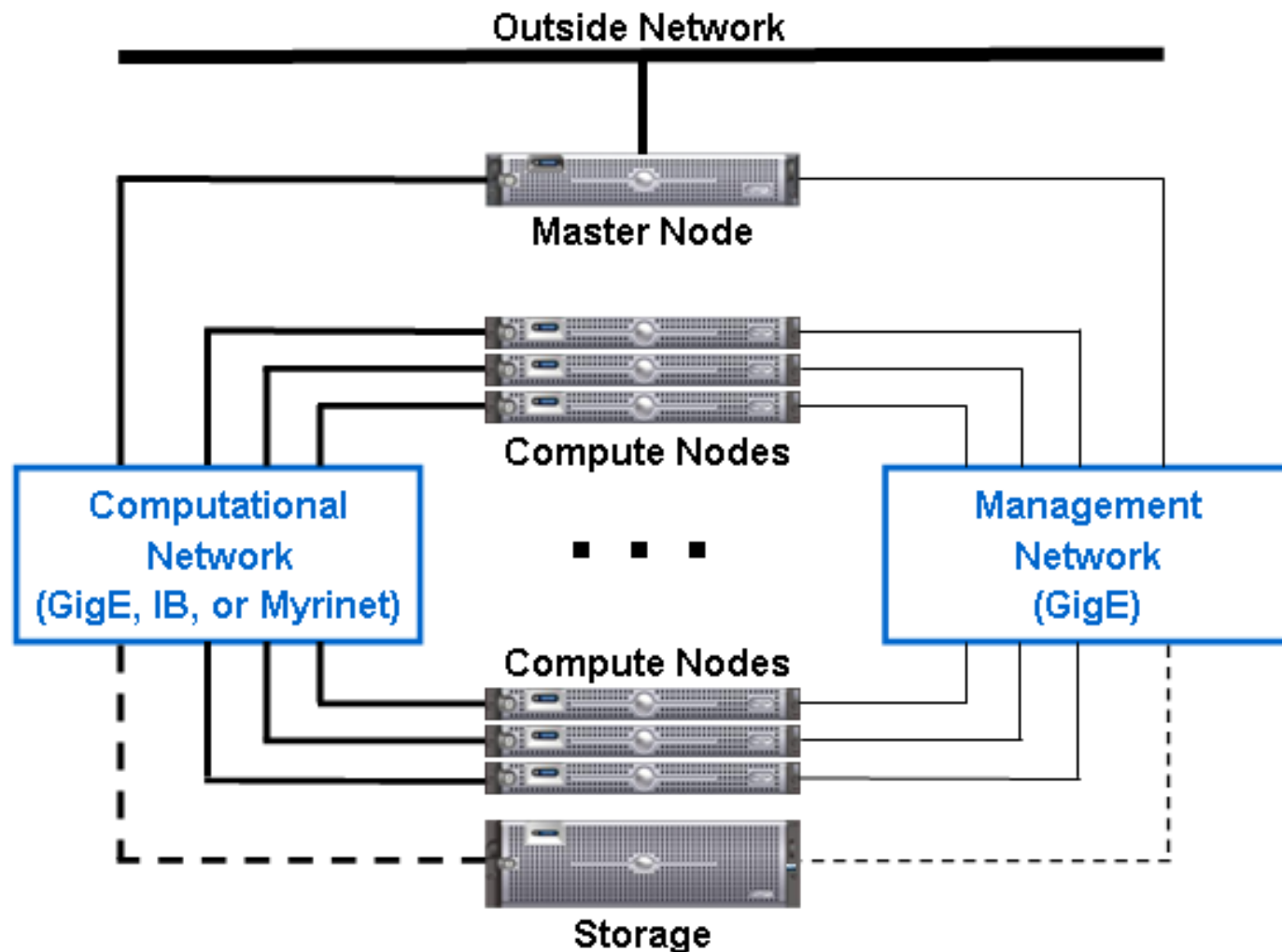
The building blocks of our computer clusters are:

- Compute **nodes**.
- Network **switches**.
- Lots of **disk*** storage



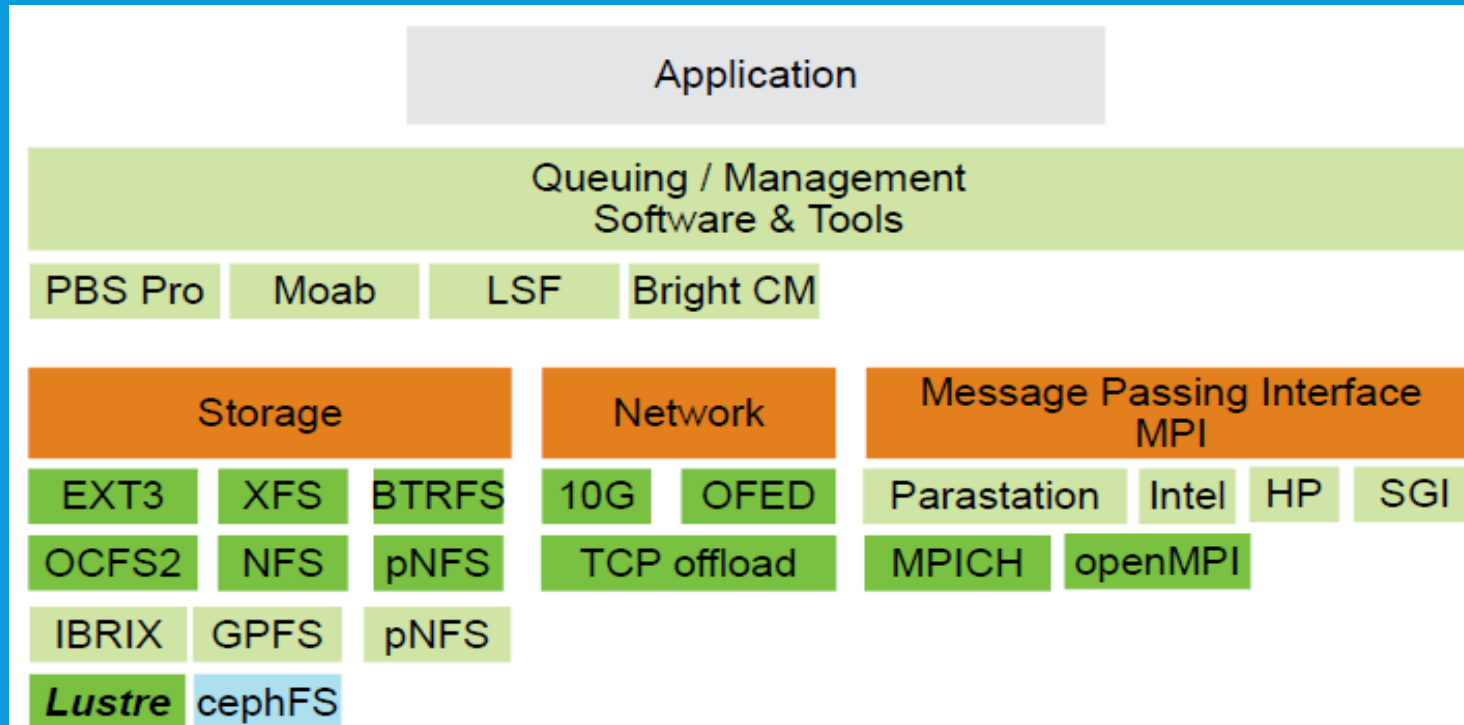
 **nexsan**
TECHNOLOGIES

COMPUTE CLUSTER ARCHITECTURE



COMPUTE CLUSTER ARCHITECTURE

Software Stack:



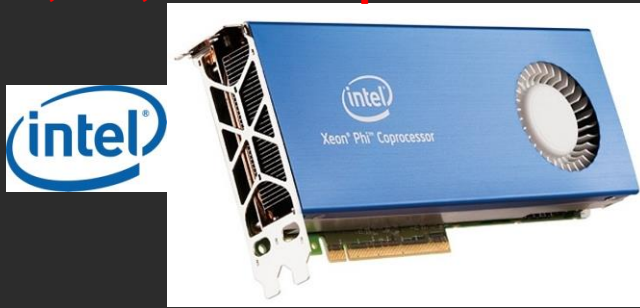
Pick Distro – Linux based (usually Enterprise class)

Hardware

A TYPICAL COMPUTE NODE

Intel Xeon Phi Coprocessor

1,200,000 MFlops



NVIDIA Tesla K20X GPU

1,310,000 MFlops



SONY



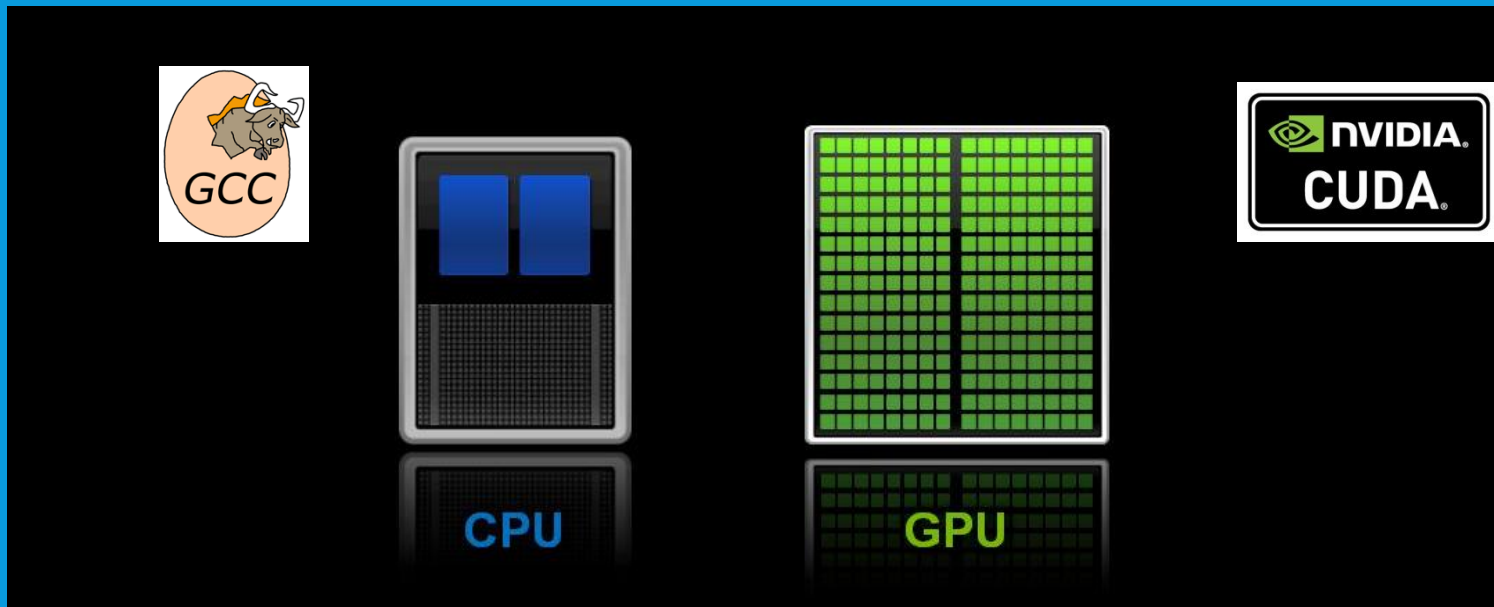
20,000 MFlops



200,000 MFlops



DIFFERENCE BETWEEN A CPU & GPU



“GPUs are optimized for taking huge batches of data and performing the same operation over and over very quickly, unlike PC microprocessors, which tend to skip all over the place.”

– Nathan Brookwood (Principal Analyst Insight64)

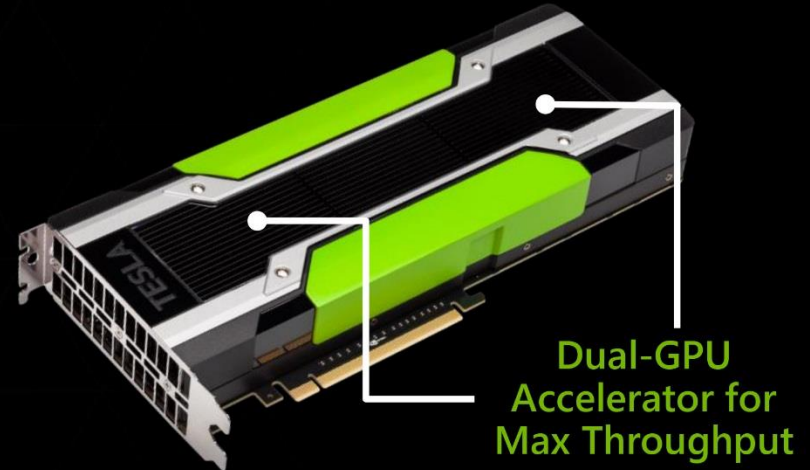
“The combination of a CPU with a GPU can deliver the best value of system performance, price, and power.”

– Kevin Krewell (Senior editor Microprocessor Report)

NEWER GPU – K40, K80

TESLA K80

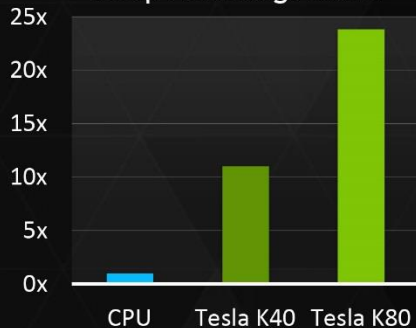
WORLD'S FASTEST ACCELERATOR
FOR DATA ANALYTICS AND
SCIENTIFIC COMPUTING



2x Faster

2.9 TF | 4992 Cores | 480 GB/s

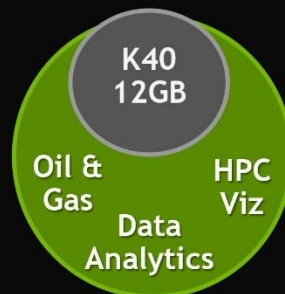
Deep Learning: Caffe



Double the Memory

Designed for Big Data Apps

24GB



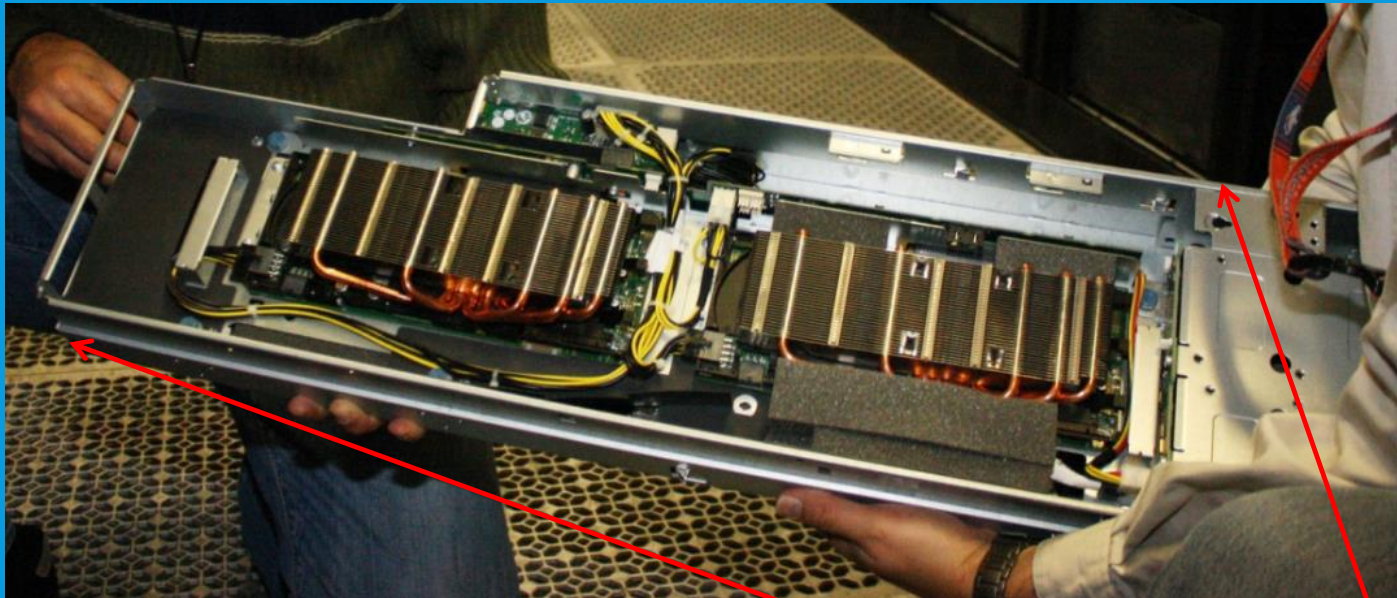
Maximum Performance

Dynamically Maximize Perf for
Every Application




GPU Boost

GPU: GRAPHICS PROCESSING UNIT(S)



INTEL PHI CO-PROCESSOR



1997: THE FIRST INTEL® TERAFLUP COMPUTER consisted of:

9,298 INTEL PROCESSORS

and occupied:


72 SERVER CABINETS
288 sq ft

THE INTEL® XEON® PHI™ COPROCESSOR will provide:

1 TERAFLUP OF PERFORMANCE

and occupy:

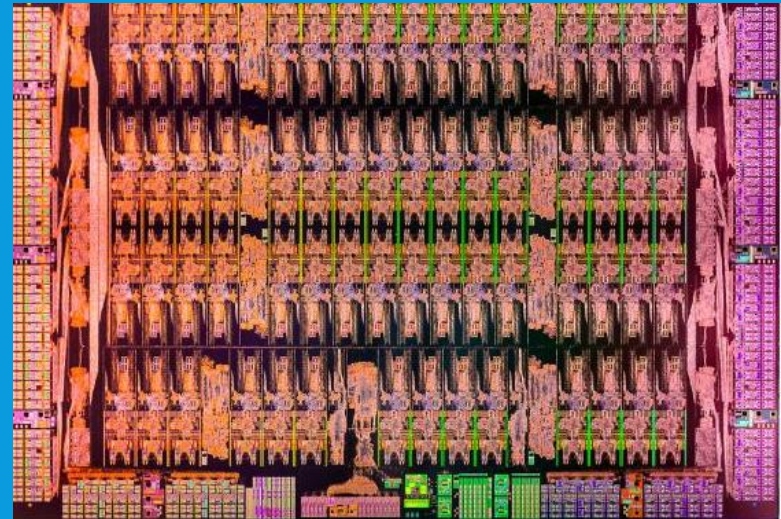
1 PCIe SLOT



Graphic courtesy of Intel Corporation

- Intel's Phi co-processor is well-suited for workloads that are memory-bandwidth bound, such as Lattice QCD and memory-capacity bound, such as ray-tracing.
- Each coprocessor contains up to 61 cores, 244 threads and 16GB of GDDR5 memory (352 GB/s bandwidth).
- The coprocessor appears as an independent server and can run Scientific Linux while consuming as low as 225 Watts. A typical CPU-based server consumes about 600 Watts.

AKA MIC = MANY INTEGRATED CORE



NETWORKING: BANDWIDTH V/S LATENCY

When selecting network switches for supercomputers we have to consider two key factors: Bandwidth and Latency . . . and price at times since some high speed switches can be prohibitively expensive.



How much can you carry?



How fast can you carry it?

INTERCONNECT SWITCHES

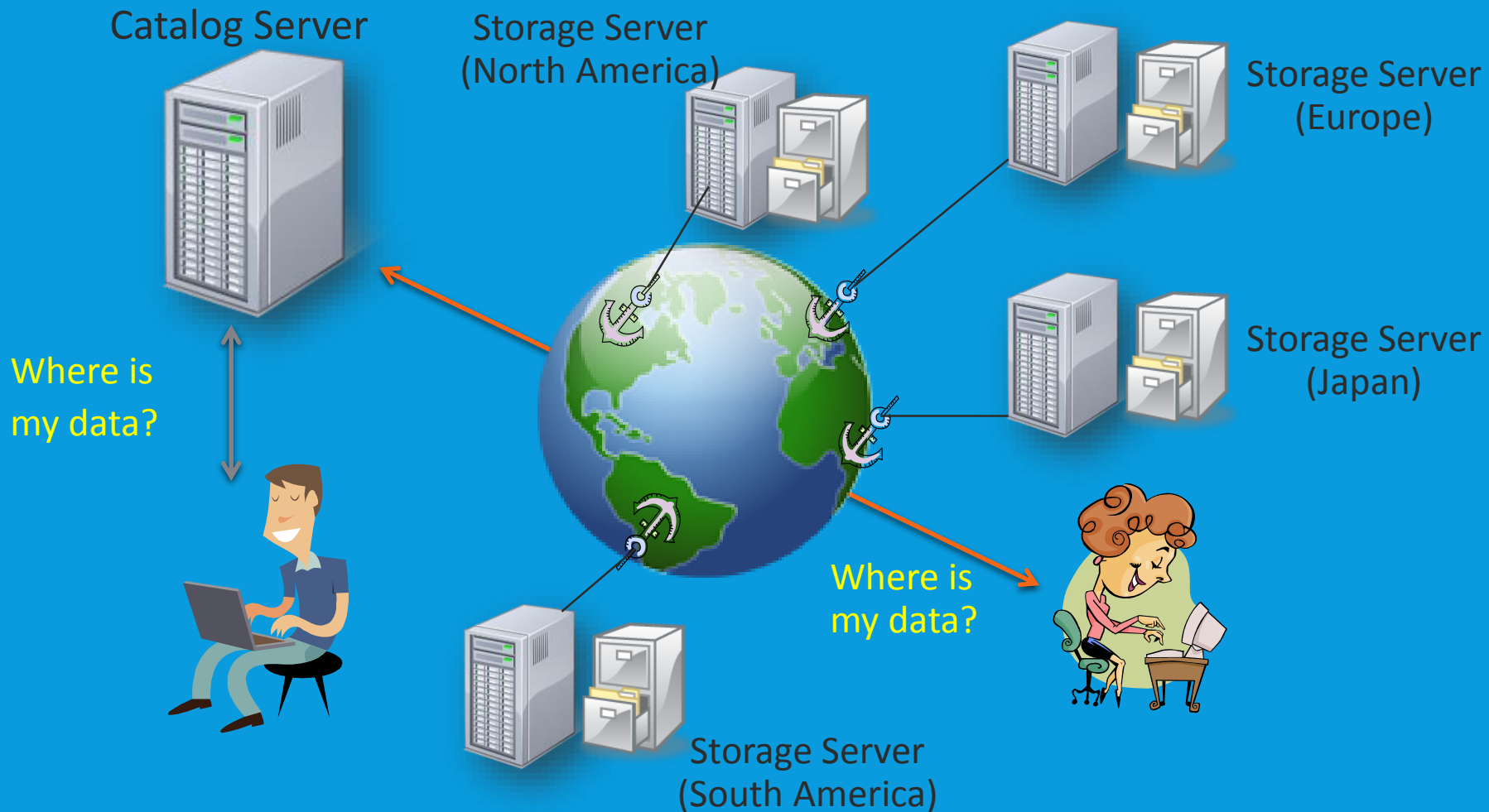


This refrigerator size network switch built by Sun Microsystems consists of 3,456 ports and is capable of transferring 14 TBytes/second which is about 3000 DVDs worth of data in one second.



We use the smaller version of this switch on our Fermilab supercomputers.

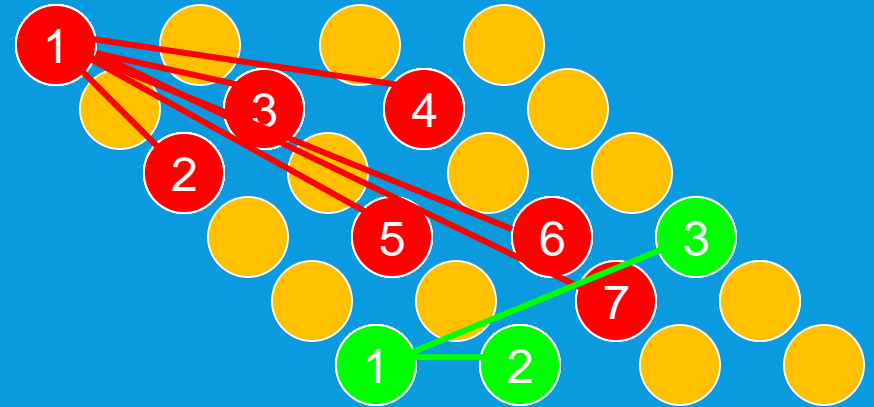
CLUSTERED STORAGE



MESSAGE PASSING INTERFACE

MPI is a language-independent communications protocol used to program parallel computers.

MPI's goals are high performance, scalability, and portability.



SIMPLE PARALLEL CODE - AN EXAMPLE

Serial to Parallel

Serial Code

Count and print the total number of characters in a book.



Parallel Code run on two (or more) computers

Odd / Even rank computer counts and prints the total number of characters in odd / even numbered chapters.



Memory



CPU



Memory



CPU

MPI Rank 1
Counts characters in odd chapters



Memory



CPU

MPI Rank 2
Counts characters in even chapters

Half the run time compared to serial code !!

MANAGING SUPERCOMPUTERS

- Biggest challenge: A job on the supercomputer will run at the speed of the slowest component.

• 2.5GHz dual-core Intel Core i5 Turbo Boost up to 3.1GHz	→ 12,400 MFlops
• 4GB of 1600MHz memory	→ 12,800 MBytes/s
• 500GB 5400-rpm Serial ATA hard disk drive	→ 3000 MBytes/s [single transfer]

The disk is 4x times slower than the CPU !!



COMPUTING FACILITIES

Feynman Computing Center



Grid Computing Center



Lattice Computing Center



OUR ESTEEMED USERS

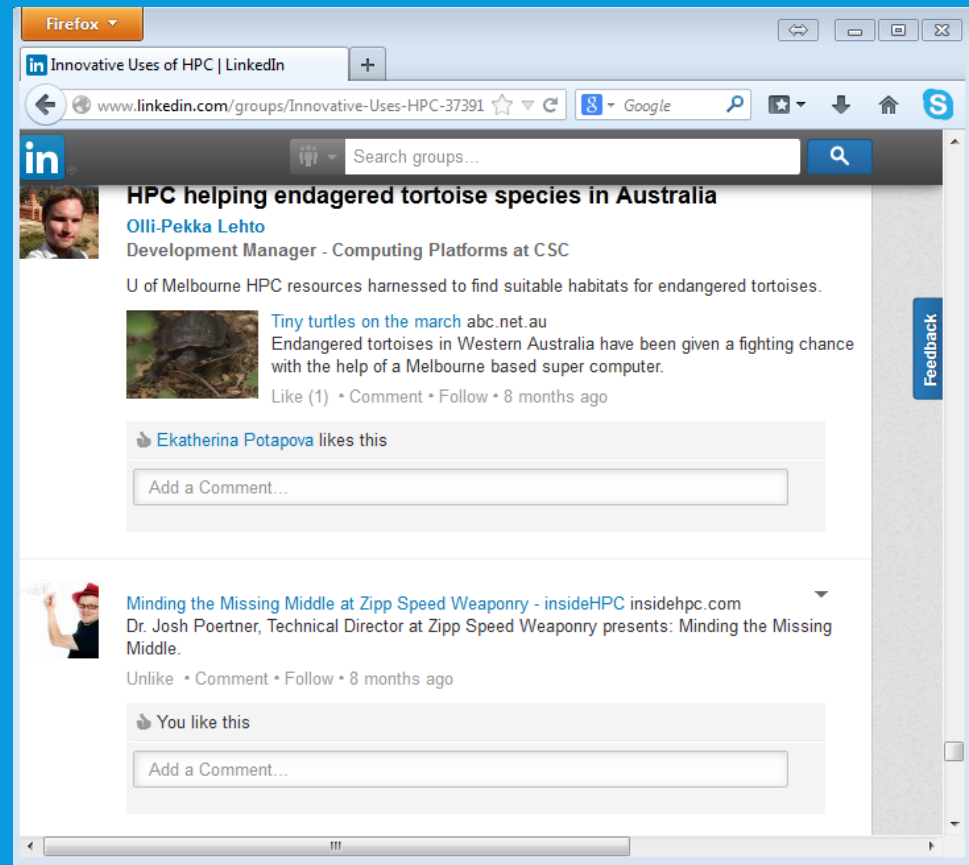
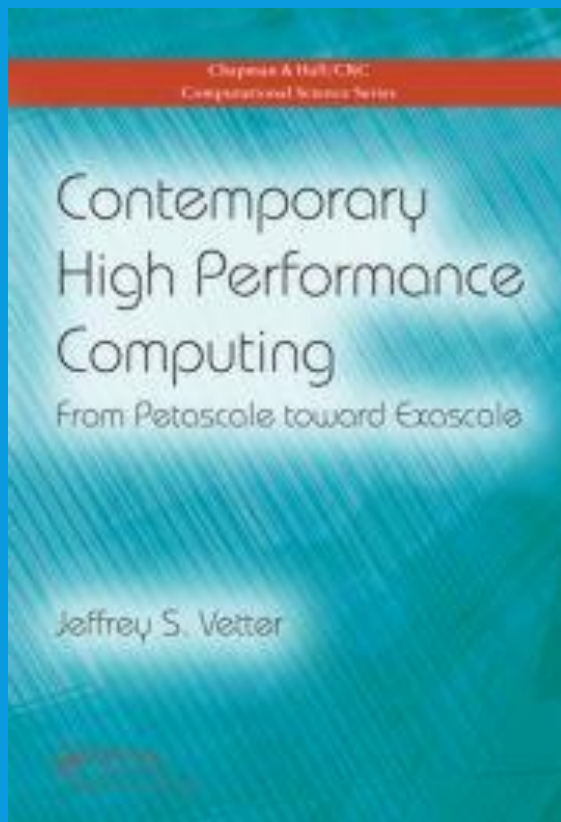


CONCLUSION

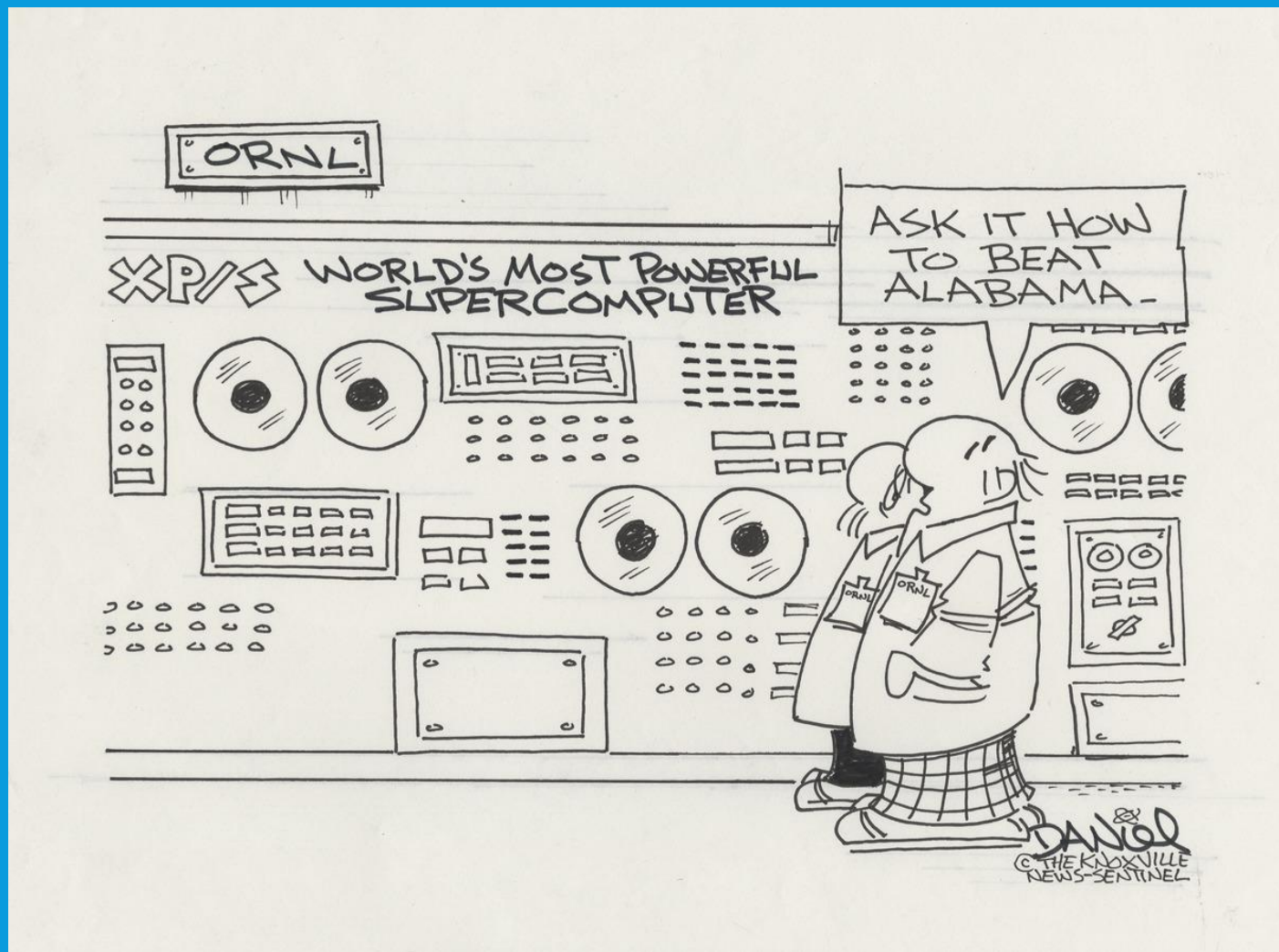
It is an exciting time to be in the field of computing which is at it's peak in terms of potential, available hardware and software options and the variety of research that can be conducted using the computing power provided by the world's fastest custom or purpose built supercomputers.

EXTRA SLIDES

FURTHER READING.....



ON A LIGHTER NOTE.....



ON A LIGHTER NOTE.....

